## Domain 3:

### Obtaining Factual Information From Data Bases

Factual data bases consist of structured knowledge that is acquired, processed, stored, and disseminated through automated electronic systems. Factual data base systems are "fact providers." They differ from bibliographic data bases, which are "fact locators," pointing to information found elsewhere. The differences between factual data bases and bibliographic data base systems are often substantial, including the methods used to construct the two types of files, the safeguards needed in choosing their content, the requirement for rigorous assessment of quality, and the desirability of repeated updating as new information replaces old. Large-scale factual data bases in computers are relatively new; the explosion of medical information and the technology to deal with it have come together only in the last 15 years.

The factual data base applications related to NLM's mission fall into three general classes: data bases for the protection of the public health and the environment, data bases providing information of special interest to research scientists in biomedicine, and data bases linked in some fashion to the provision of health care and the practice of the health professions. In addition, these data bases are often assembled to support expert systems or modeling systems. User-cordial linkages from factual data bases to such systems are becoming increasingly important resources in several areas of biomedicine.[33]

### Data Bases For The Protection Of Public Health

The Library's commitment to factual data bases for the protection of the public health and the environment is exemplified by the TOXNET online toxicology information system.[34] The data bases in that system describe the effects of chemical substances on humans, other biological systems, and the environment. The number of chemicals that could pose a public health hazard is relatively small (less than 10,000) compared with the total number of known chemicals (over 7.5 million.) Consequently, substantial economies can be achieved, nationally and internationally, by collecting authoritative descriptions of the biological and environmental effects of hazardous chemicals in one central data base or in a few well-coordinated data base building and maintenance efforts. Dissemination of the resulting body of data can then take place through various public and private sector channels. To avoid duplication of expensive efforts, collaboration with other federal and state agencies through the sharing of funding and other resources—including intellectual resources—is particularly important.

Clearly NLM will not—and should not—be the sole provider of information about hazardous chemicals. However, the existence of multiple data bases, local as well as national and international, located on different computer systems and using different query systems, makes it difficult for even an experienced information specialist to access all the relevant online services.

The liability issues identified in connection with the provision of factual data base services in general are especially important here. Risk assessment and risk management decisions made on the basis of data provided by factual data base services in toxicology can have major impacts on human health, the environment, and the economy. The responsibility the Library carries for the accuracy and currency of the data in such data bases, therefore, is substantial. Conventional methods of content review to assure accuracy and currency are based on consensus of expert panels and are effective, but slow, cumbersome, and expensive. While the Library has made a good beginning in employing modern electronic measures—such as computer conferencing—for the peer review of data base content, technologies are available to enhance those methods even further.

Computer-based modeling that attempts to predict the biological activities of chemicals based on the known activities of structurally related chemicals can play an important role in developing data for risk assessment, synthesis of new pharmaceutical or agricultural products, and reduction in the number of animals needed for biological research and testing.[35] There is a need, therefore, to foster the development and operation of such modeling systems by ensuring that the content and structure of applicable data bases are suitably organized.

Increasingly, the Library's factual data bases in this category will be used by persons responsible for responding to emergencies involving hazardous chemicals. Under such conditions, emergency responders will need selected, simplified, or summarized effects and treatment data. It is likely that some responders will lack experience in computer searching and data manipulation. Therefore, highly user-cordial information systems will be required by all those using data for human and environmental protection from hazardous chemicals.

### Biomedical Research Data Bases

The field of molecular biology is opening the door to an era of unprecedented understanding and control of life processes. Automated methods are now available to analyze and modify biologically important macromolecules. The effects of this research are already evident in clinical medicine. The prenatal diagnosis of blood disorders, such as thalassemia and sickle cell disease, has only recently been made possible through newly acquired genetic knowledge and the production of therapeutic agents, such as interferons and interleukins, depends on DNA and protein sequence information assembled in accessible data bases.

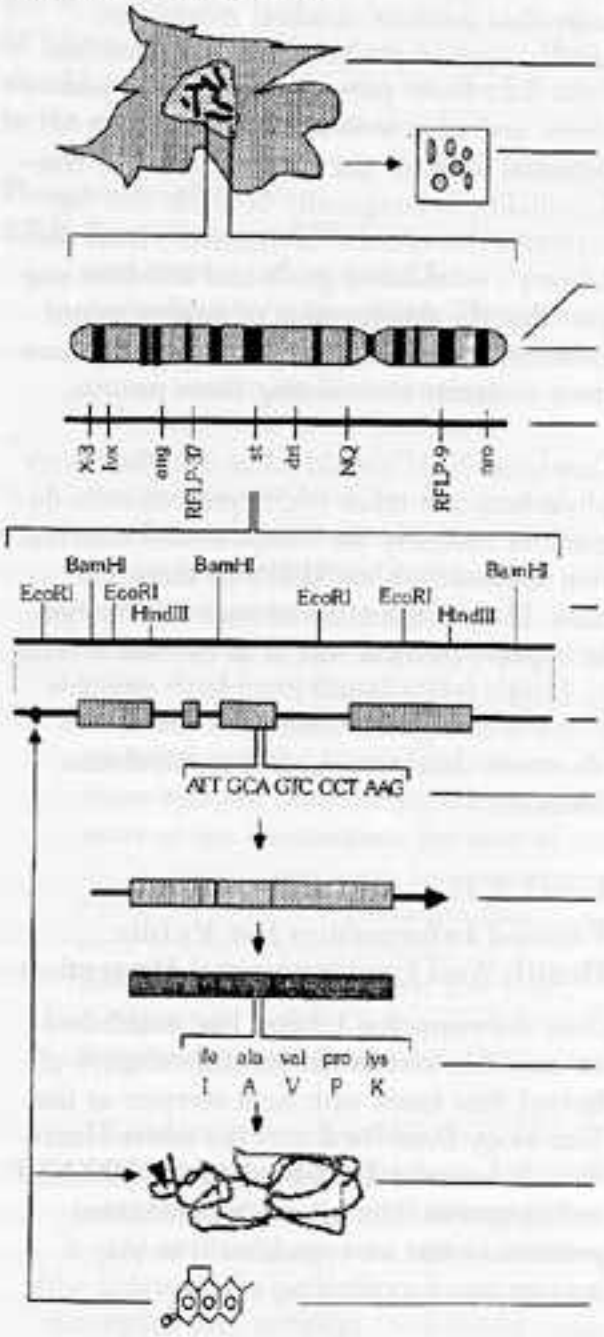
In molecular biology, factual data bases have become a necessity for scientific research.[36] Because of the complexity of biological systems (for example, the human genome is thought to be made up of 3 billion DNA base-pairs) basic research in the life sciences is increasingly dependent on automated tools to store and manipulate the large bodies of data describing the structure and function of important macromolecules.[37] Factual data bases have been developed to store data relating to each level of the natural hierarchy, from cells through successively smaller genetic units, to base-pair sequences.[38] However, the relatively isolated design of the various data bases contrasts sharply with cur-

rent research activities in molecular biology, where an investigator will commonly report findings involving data at the cellular, chromosomal, gene, amino acid, and DNA sequence levels within a single scientific paper. Although the critical scientific questions being asked can often be answered only by relating one biological level to another, methods for automatically suggesting links across levels are non-existent.

Currently, no organization is taking the leadership to promote keys and standards by which the information from the related research data bases can be systematically interlinked or retrieved by investigators. The full potential of the rapidly expanding information base of molecular biology will be realized only if an organization with a public mandate such as the Library's takes the lead to coordinate and link related research data bases, and make them easily accessible to the U.S. and international research community.

## Practice-linked Factual Data Bases

Health care professionals must have access to a vast and rapidly changing body of knowledge concerning the proper management of human illnesses. Traditional paper-based methods of information transfer in the health sciences are inadequate at present and will become more so in the future. Thus, providing authoritative information to help health-care professionals make decisions will become an increasingly important activity for government health agencies and professional societies.[39,40] Such information should be complemented, where appropriate, by appendiceal data files of selected pre-clinical and clinical research results. In addition to being useful repositories of facts for practitioners, factual data bases can also provide the basis for expert systems such as those developed experimentally for making diagnoses and determining treatment regimens in medicine.[41,42,43]




*Biomedical Data Bases in a Universal Hierarchy of Nature: cells—chromosomes—genes—proteins.*

Factual data bases containing all or part of individual patients' medical records are another type of practice-linked information. Such data bases present some special problems and raise serious questions about potential roles for the Library. Issues of confidentiality, heterogeneity of needs and format, and the substantial departure from the Library's established goals and activities suggest that the development of patient record systems be left with the health care organizations currently maintaining those records.

One issue NLM should address is that many physicians and other health professionals do not now routinely use computerized information sources such as NLM's in their practices. If the routine use of such information to improve medical care is to become a reality, health professionals must have available better training, education, and practice in electronic data retrieval and manipulation methods.

## Goal 3.1:
## Expand Information For Public Health And Environmental Protection

Over the years, the Library has established an excellent foundation for this category of factual data bases with such services as the Toxicology Data Bank and the newer Hazardous Substances Data Bank in the TOXNET online system. The Library's preeminent position in this area qualifies it to play a leading and coordinating role, working with other government agencies and the private sector.

The Library should accommodate public health needs as far as possible in data base content organization and in developing computer methodologies, including the use of artificial intelligence. Special attention should be given to tailoring data representation and retrieval to emergency and occupational safety applications. For such activities, NLM should be provided with the required resources—including guidance about requirements and close cooperation in implementation—by those agencies specifically charged with chemical emergency response and occupational health and safety. The Library should then actively try to share the resulting access and delivery methods with other interested agencies at all levels of government, including international organizations.

### Recommendations

3.1.1. Continue the maintenance and enhancement of the Hazardous Substances Data Bank and the other factual data bases now provided through the TOXNET system. Ancillary factual data bases of particular utility for occupational safety and health should be acquired from other sources (both nationally and internationally) or built by the Library when required. Wherever possible, file building and enhancement costs should be shared with other federal agencies that have specific mandates in these areas.

3.1.2. Continue the mutually useful collaboration with the ATSDR (Agency for Toxic Substances and Disease Registry) on the information requirements of CERCLA (Comprehensive Environmental Response, Compensation, and Liability Act, or Superfund) as reauthorized in 1986. The focus should be on preparation of extensive profiles of selected hazardous chemicals. NLM's work on the profiles should be compensated by CERCLA through ATSDR.

3.1.3. Take a national coordinating role for Federal and State activities in building and maintaining factual data bases on the biological and environmental effects of hazardous chemicals. Such coordination should lead to efficien-

cies making the resulting products more widely useful and cost effective.

3.1.4. Continue to develop gateway systems to facilitate access to and use of data about hazardous chemicals located in different public and private systems. Because such systems will be used in chemical emergency situations, partial support for their development should come through CERCLA.

3.1.5. Continue and increase its efforts to ensure the quality of its factual data bases through ongoing content review by subject experts. The Library should also research and develop ways of further improving the efficiency of this process, perhaps using electronic methods to eliminate the need for panel meetings altogether.

3.1.6. Support national and international modeling and analytical activities particularly as they pertain to relating biological activities to chemical structures. Toward that end, NLM should maintain relevant data bases and user-cordial gateways to existing modeling activities.

## Goal 3.2:
## Establish Information Services and Linkages For Biotechnology Information

A singular and immediate window of opportunity exists for the Library in the area of molecular biology information. Because of new automated laboratory methods, genetic and biochemical data are accumulating far faster than they can be assimilated into the scientific literature. The problems of scientific research in biotechnology are increasingly problems of information science. By applying its expertise in computer technologies to the work of understanding the structure and function of living cells on a molecular level, NLM can assist and hasten the Nation's entry into a remarkable new age of knowledge in the biological sciences. This should remain a high priority for the Library in the coming two decades.

### Recommendations

3.2.1. Immediately establish an intramural and extramural program for biotechnology information. The intramural component should be a National Center for Biotechnology Information, to serve both as a repository and distribution center for this growing body of knowledge and as a laboratory for developing new information analysis and communications tools essential to continued advancement in this field. The program should emphasize collaboration between computer and information scientists and the biomedical researchers who are both the producers and users of the information. Because of the technical complexity in this scientific area and the expectation that data production will increase by a thousand times in the next five years, a major new activity is required. Specifically, the Library should:

- Conduct research in the areas of molecular biology data base representation, retrieval-linkages, and modeling systems while examining analytical interfaces based on algorithms, graphics, and expert systems.

- Provide repository, directory, and distribution services in the areas of data collection and quality control, as well as online data delivery through linked regional centers and distributed data base subsets.

- Develop and implement training workshops, information clearinghouse activities, and documentation programs.

3.2.2. Sponsor meetings that include a broad representation of the scientists responsible for designing and maintaining of current research-oriented, genetic factual data bases. The purpose of those meetings will be to develop a consensus regarding the best methods for information sharing and retrieval from related molecular biology data bases.

## Goal 3.3:
## Support The Development Of Medical Practice-linked Data Bases

As the nation's health care practitioners become more familiar with the inherent advantages of computer-based data systems and more willing to use them, practice-linked factual data bases can be expected to become more numerous. Now is the time to begin a coordinated approach to designing and implementing those systems The work should be based on standards that make optimal use of the emerging technology and of the information contained in the data bases themselves.

The Library is best positioned to take a principal, coordinating role in this developing area because of its acknowledged leadership in the area of biomedical information and communications.[44] The Library's role should be to provide the system design team and technical expertise for other organizations that would be responsible for the content of the data bases. Because the program will place additional responsibilities on the Library without diminishing its traditional mandate, funding should be sought from new appropriations rather than reprogramming existing resources. Wherever feasible, program costs should be shared with the organizations responsible for data base content.

An important component of increasing the usefulness of such information sources will be research into the design and construction of full-text, natural language retrieval systems with visible links among related data bases.

### Recommendations

3.3.1. Establish an intramural program capable of developing practice-linked data bases in collaboration with public and private health care agencies, including other institutes of the NIH. The program should promote factual data base standards; for example, the Unified Medical Language System.

Once established, the program's services should be actively promoted within the NIH and to the academic medical community. The Library should also develop models for sharing the development costs of new factual data bases and for ongoing cost reimbursement through licensing agreements with public agencies and private vendors.

3.3.2. Develop specialized pseudo-English or menu-driven interfaces for certain factual data bases. Initially, one practice-linked and one biomedical research data base should be chosen. The medical data base interface may, in fact, be subsumed by work on a Unified Medical Language System, and its development costs be viewed as an integral part of that effort.

3.3.3. Signify NLM's willingness to store and make available appendiceal data files of selected published research.

### Budget

Estimates of resources needed to implement these recommendations are given in Chapter 4.